When speed matters in learning against adversarial opponents

(Extended Abstract)

Mohamed Elidrisi
Dept. of Computer Science and Engineering
University of Minnesota
Minneapolis, MN 55455
elidrisi@cs.umn.edu

Maria Gini
Dept. of Computer Science and Engineering
University of Minnesota
Minneapolis, MN 55455
gini@cs.umn.edu

ABSTRACT

We propose a novel algorithm that is able to learn and adapt to an opponent even within a limited number of interactions and against a rapidly adapting opponent. The context we use is two player normal form games. We compare the performance of an agent using our algorithm against agents using existing multiagent learning algorithms.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*Intelligent agents*

General Terms

Algorithms, Economics, Experimentation

Keywords

Multiagent Learning, Game Theory, Adaptive Learning

1. INTRODUCTION

A challenging issue in the design of intelligent agents is how to endow them with the ability to interact with other intelligent agents. Multiagent learning is primarily concerned with the problem of learning and acting in the presence of opponents. Multiagent learning has received considerable attention in the past decade from the research community, which has produced a wide range of learning agents and a set of criteria for developing them. Within the AI community, the problem has been addressed in multiple ways, either by adapting single agent reinforcement learning algorithms for multiagent settings [3], or combining policy search with knowledge of the adversarial nature of the opponent [1], or from a game theoretic perspective [4].

One of the major constraint typically assumed is that the opponent is either stationary or will converge to a stationary policy [1]. The stationarity assumption has been relaxed to some degree (e.g. [4]), but there are still critical assumptions that limit the use of learning agents in real world domains. We investigate two of those. The first relates to the need for extremely long sequences of interactions between the agents, often in the order of hundreds of thousand, before the agent learns a policy to use against the opponent. The second relates to the fact that abrupt changes in the opponent's play often require to restart the learning process.

2. A NEW ALGORITHM: FAL

We propose a novel algorithm, *Fast Adaptive Learner (FAL)*, to learn a strategy to use when playing a sequence of games against an opponent. A strategy in a repeated game is a mapping from the history of actions to a probability distribution over the actions. The key feature of our algorithm is the ability to learn in a limited number of interactions and to detect and adapt to potentially fast changes in the opponent's strategy.

The algorithm, at a high level, uses two models:

1. a *Predictive Model* which makes a prediction about the opponent's next action. The predictive model has to be online in nature with some decay function over the history of interactions. It also has to view the interactions as a sequential prediction problem not as independent predictions and detect abrupt changes in the interactions.
2. a *Reasoning Model* which chooses a suitable best response accordingly. The reasoning model needs a belief model of whether the opponent is cooperative or competitive and the ability to explore if the opponent is teachable. It should also be able to measure the success of the predictive model in addition to maintaining a target average reward as a safety value.

There is a large class of models and methods that can be used for both parts of the algorithm. We made specific choices for the models used in our experiments, but we are not limited to the models we used. It is important to note that the memory of the predictive model limits the target class of opponents FAL can adapt to.

3. EXPERIMENTAL RESULTS

We have instantiated FAL's predictive model with ELPH and its reasoning model with Godfather-Future.

ELPH [2] is an an online predictive algorithm that learns to predict from short sequences. ELPH keeps a hypotheses space with the patterns observed and predictions sets that are updated constantly and pruned using entropy.

Appears in: *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)*, Conitzer, Winikoff, Padgham, and van der Hoek (eds.), 4-8 June 2012, Valencia, Spain.

3.0,3.0	0.0,5.0
5.0,0.0	1.0,1.0

Table 1: Prisoner's Dilemma game matrix

	Q1	WOLF-PHC	FAL	God Father	Bully
Q1	1.7,1.7	1.7,1.7	2.2,2.2	2.4,2.4	0.9,1.4
WOLF-PHC		1.9,1.9	2.2,2.2	2.4,2.4	0.9,1.4
FAL			3.0,3.0	3.0,3.0	1.0,1.0
GodFather				3.0,3.0	1.0,1.0
Bully					1.0,1.0

Table 2: Average pairwise payoffs after playing 100 rounds of Prisoner's Dilemma. Results are from 1000 runs.

To explain Godfather-Future, we need a few concepts from game theory. A *security value* is the strategy that maximizes the player's own minimum payoff. A *targetable pair* is any pair of deterministic strategies in the game with the property that it yields a reward for the player higher than its security value. The Godfather-Future strategy computes a targetable pair of actions that leads to higher reward than its security value. The original Godfather [3] plays its part of the targetable pair if the opponent played its half of the targetable pair in the last interaction, Godfather-Future plays its part if the opponent is predicted to play its part in the next interaction.

Experiment 1. We compared experimentally the performance of different learning algorithms, using two-player repeated normal form games. The results in Table 2 show the outcomes of playing Prisoners Dilemma for 100 iterations. We repeated each of the 100 iterations 1000 times to reduce noise. The performance of FAL is compared against a set of algorithms and strategies from the literature, specifically Q-Learning, WOLF-PHC [1], Bully, and Godfather [3].

In Prisoner's Dilemma, shown in Table 1, the dominant strategy is to defect (D) and receive a reward of 1.0. Cooperating (C) would lead to a higher outcome of 3.0 but with the added risk of getting 0 if the opponent decided to betray.

From Table 2 it is clear that FAL and Godfather are the best performing methods across the board. Q-Learning and WOLF-PHC were among the worst especially against a stationary policy like Bully. It is important to note that Q-Learning and WOLF-PHC will perform as well as the other agents in longer sequences of games but our goal in this research is to analyze short term performance.

Experiment 2. In Experiment 1 we have shown that FAL is able to learn faster and achieve better results than Q-Learning, WOLF-PHC, and Bully. However, the performance of Godfather and FAL were almost identical in many scenarios. In order to show the ability of FAL to adapt rapidly we present now results against an opponent that changes its strategy after some period of time. Detecting the change and adapting to it is the real advantage that we are aiming at achieving in this work.

We use as opponent an agent we call *Switch*. The agent starts by following the classical Godfather strategy until it reaches stage 40 of the game. After that, Switch follows a deterministic repeated sequence of actions C, D, C, C, D, C indefinitely. This agent is designed to be deterministic and predictable with a bounded memory.

In this experiment, Switch played a sequence of 100 games against FAL, Godfather, and WOLF-PHC. Figure 1 shows the average reward over time for the 3 agents against Switch. Positive values imply that Switch is getting more reward, 0 are ties, and negative values are the others. It is evident in the graph that FAL is the best performing agent. FAL is able to detect and adapt in less than 20 games to the opponent's policy changes while the rest of the agents were not able to detect it until end of sequence at game 100.

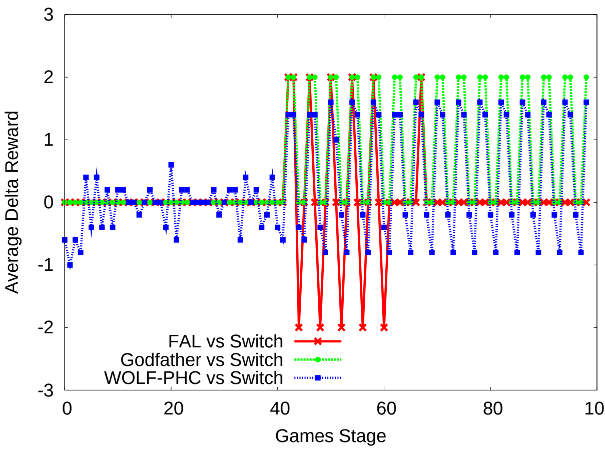

Figure 1: Average delta reward for the 3 Agents vs. Switch agent.

4. CONCLUSIONS AND FUTURE WORK

The goal of this work is to motivate and introduce the need for new requirements on multiagent learning algorithms, specifically to create agents that learn after playing a limited number of games against an opponent and that are capable of adapting to sudden and frequent changes in the opponent strategy. We proposed a new algorithm, FAL, and demonstrated experimentally that FAL outperforms agents using other learning methods in the Prisoner's Dilemma and against an abrupt policy changing opponent. Future work will be directed at examining theoretical properties of FAL, applying it to a larger class of games, and expanding the algorithm to play against more than one opponents.

5. REFERENCES

[1] M. Bowling and M. Veloso. Multiagent learning using a variable learning rate. *Artificial Intelligence*, 136(2):215–250, 2002.

[2] S. Jensen, D. Boley, M. Gini, and P. Schrater. Non-stationary policy learning in 2-player zero sum games. In *Proc. Nat'l Conf. on Artificial Intelligence*, pages 789–794. AAAI Press, 2005.

[3] M. Littman and P. Stone. Leading best-response strategies in repeated games. In *Int'l Joint Conf. on Artificial Intelligence Workshop on Economic Agents, Models, and Mechanisms*, 2001.

[4] R. Powers, Y. Shoham, and T. Vu. A general criterion and an algorithmic framework for learning in multi-agent systems. *Machine Learning*, 67(1):45–76, 2007.